Pearson's correlation coefficient between X and Y for sample size n

definition

$$-1 \le r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \le 1$$

Where

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

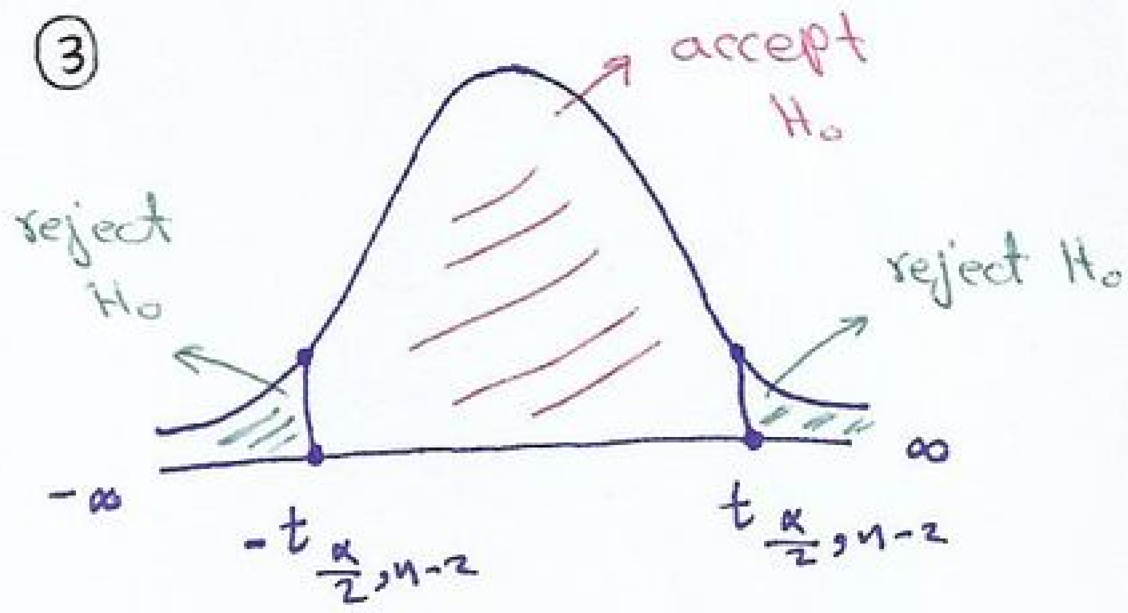
Hypotheses testing

① $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

② test statistics

$$t_{T.S.} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{v=n-2}$$

③



④ we reject H_0 if:

$t_{T.S.} < -t_{\frac{\alpha}{2}, n-2}$

or

$t_{T.S.} > t_{\frac{\alpha}{2}, n-2}$

otherwise, we will accept H_0

or

We reject H_0 if:

$p\text{-value} \le \alpha$

where

$p\text{-value} = 2P(T > |t_{T.S.}|)$

Some notes:

① Pearson's correlation coefficient for stady the streight line/ linear relationship between X and Y

② $\rho \longrightarrow$ correlat. for population

$r \longrightarrow$ Correlation for sample

③ (i) Sign of r

$\longrightarrow$ + $\longrightarrow$ Positive direction/Correlation i.e.

increase - decrease X $\longleftrightarrow$ increase - decrease Y

$\longrightarrow$ - $\longrightarrow$ negative direction/ correlation i.e.

increase - decrease X $\longleftrightarrow$ decrease - increase Y

(ii) Value of r

$\longrightarrow |r| = 0 \longrightarrow$ no correlation/ no linear relationship

$\longrightarrow 0 < |r| \leq .25 \longrightarrow$ weak

$\longrightarrow .25 < |r| < .75 \longrightarrow$ moderate

$\longrightarrow .75 \leq |r| \longrightarrow$ strong

"Simple linear regression"

dependent/response variable

independent/predictor/explanatory variable

where

$Y_i \sim N\left(E(Y_i) = \beta_0 + \beta_1 X_i , \; Var(Y_i) = \sigma^2 \right)$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \; i = 1, 2, \ldots, n$$

Parameter

(constant/intercept)

Parameter

(coefficient/slope)

error

where

$\varepsilon_i \sim N\left(E(\varepsilon_i) = 0 , \; Var(\varepsilon_i) = \sigma^2 \right)$

and

$Cov(\varepsilon_i, \varepsilon_j) = 0, \; \forall i \neq j$

"fitted regression line/equation"

or

"Prediction equation"

Fitted Value

independent/predictor variable

$$\hat{Y}_i = b_0 + b_1 X_i$$

constant/intercept

coefficient/slope

① $b_0 = \begin{cases} \bar{Y} - b\bar{X} \\ \sum_{i=1}^{n} L_i Y_i \; ; \; L_i = \frac{1}{n} - \frac{(X_i - \bar{X})}{S_{XX}} \end{cases}$

② $E(b_0) = \beta_0$ i.e b_0 is unbiased estimate of β_0 ;

$Var(b_0) = MSE\left(\frac{1}{n} + \frac{(\bar{X})^2}{S_{XX}}\right)$;

standard error of b_0 is $S.E.(b_0) = \sqrt{Var(b_0)}$

③ The sampling distribution of b_0:

$T_0 = \frac{b_0 - \beta_0}{S.E.(b_0)} \sim t_{n-2}$

④ $100(1-\alpha)\%$ confidence interval of β_0:

$\beta_0 \in b_0 \pm t_{\frac{\alpha}{2}, n-2} \; S.E.(b_0)$

① $b_1 = \begin{cases} \frac{S_{XY}}{S_{XX}} = r\sqrt{\frac{S_{YY}}{S_{XX}}} \\ \sum_{i=1}^{n} K_i Y_i \; ; \; K_i = \frac{(X_i - \bar{X})}{S_{XX}} \end{cases}$

② $E(b_1) = \beta_1$ i.e b_1 is unbiased estimate of β_1 ;

$Var(b_1) = \frac{MSE}{S_{XX}}$;

standard error of b_1 is $S.E.(b_1) = \sqrt{Var(b_1)}$

③ The sampling distribution of b_1:

$T_1 = \frac{b_1 - \beta_1}{S.E.(b_1)} \sim t_{n-2}$

④ $100(1-\alpha)\%$ confidence interval of β_1:

$\beta_1 \in b_1 \pm t_{\frac{\alpha}{2}, n-2} \; S.E.(b_1)$

⊛ b_0 and b_1 are estimate of β_0 and β_1, respectively,

by

using

the least squares method / minimization procedure

Some notes:

① The error/residual $e_i = Y_i - \hat{Y}_i$, then:

ⓘ sum of squared error:

$$SSE = \sum_{i=1}^{n} e_i^2$$

ⓘⓘ $\sum_{i=1}^{n} e_i = 0$

ⓘⓘⓘ mean squared error:

$$MSE = \hat{\sigma}^2 = S^2 = \frac{SSE}{n-2};$$

and $E(S^2) = \sigma^2$ i.e S^2 is unbiased estimate of σ^2

② $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

③ The sum of weighted residuals:

ⓘ by X : $\sum_{i=1}^{n} X_i e_i = 0$

ⓘⓘ by Y : $\sum_{i=1}^{n} Y_i e_i = 0$

④ regression line through $(\bar{X}, \bar{Y})$

⑤ for K_i : ⓘ $\sum_{i=1}^{n} K_i = 0$

ⓘⓘ $\sum_{i=1}^{n} K_i X_i = 1$

ⓘⓘⓘ $\sum_{i=1}^{n} K_i^2 = \frac{1}{S_{xx}}$

for L_i : ⓘ $\sum_{i=1}^{n} L_i = 1$

ⓘⓘ $\sum_{i=1}^{n} L_i X_i = 0$

ⓘⓘⓘ $\sum_{i=1}^{n} L_i^2 = \frac{1}{n} - \frac{(\bar{X})^2}{S_{xx}}$

⑥ $(SSToT = S_{yy}) = SSE + SSR$

total sum of squared error

sum of squared regression $= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

⑦ The coefficient of determination

$$0 \leq R^2 = \frac{SSR}{SSToT} = 1 - \frac{SSE}{SSToT} \leq 1$$

where:

ⓘ $100R^2\%$ of total variation in Y is due to X.

or

(the least squares regression/regression line) explains $100R^2\%$ of the variation in Y.

ⓘⓘ The simple correlation coefficient $r = \sqrt{R^2}$ and the sign of it will taken from b_1.